PROCESADOR ACUSTICO DE PALABRAS BASADO EN LA TECNICA DE "WORD SPOTTING" PARA EL RECONOCIMIENTO DEL HABLA CONTINUA.

J. Salavedra , E. Lleida , J. B. Mariño , C. Nadeu
Departament de Teoria del Senyal i Comunicacions
E.T.S.E.Telecomunicació. U.P.C. 08080 Barcelona

ABSTRACT

This paper describes the acoustic processor of a Spanish Continuous Speech Recognition System based on HMM Demisyllable units. The acoustic processor produces a lattice of word hypotheses suitable to be parsed by a linguistic analyzer and it is based on a spotting algorithm. This spotting algorithm has three inputs : The unknown utterance, the HMM of the demisyllables or references and the lexical knowledge in terms of a finite state network . In this paper , the authors describe a less cost method for speaker independent continuous speech Word Spotting.This spotting algorithm is a modified version of the One-Step Viterbi algorithm with multiple hypothesis. The implemented acoustic processor has been tested using telephonic numbers and integers numbers from 0 to 1000 in a speaker-independent context : Excellent results have been obtained that have demostrated the efficiency of the spotting algorithm and the performance of the demisyllable as recognition unit.

1 . INTRODUCCION

En el ámbito del reconocimiento de voz por computador, se advierte una tendencia bastante generalizada a centrar las lineas de investigación hacia la obtención de sistemas de comprensión del habla en un contexto de independencia del locutor y con un vocabulario de palabras considerable. A partir de esto, resulta fácil preveer que, si bien existen diversos tipos de aproximaciones y prototipos, parece más que probable que los futuros sistemas de comprensión del habla cuenten con un módulo para detectar palabras. La técnica de "Word Spotting" desarrollada en este trabajo encuentra su justificación en este contexto.

Este artículo se estructura en las siguientes secciones : La sección 2 describe las características principales del sistema. La sección 3 trata sobre el primer nivel de localización considerado, la semisílaba (SM), y los algoritmos de localización de SMs.. La sección 4 considera el conocimiento léxico y los algoritmos de localización de palabras. La sección 5 muestra los resultados experimentales. Y, finalmente, en la sección 6 se presenta una breve conclusión.

2 . SINTESIS Y ARQUITECTURA DEL SISTEMA

La arquitectura básica del sistema se representa en la Figura 1, según un diagrama de bloques . El proceso de reconocimiento consta de dos etapas básicas : La etapa de

Este trabajo ha estado financiado por la PRONTIC nº 105/88

entrenamiento y la de reconocimiento. Estas se hallan explícitamente identificadas en el esquema de la figura 1.

Este trabajo (procesado acústico) se centra, básicamente, en la implementación del corazón del sistema : El algoritmo de "Spotting". Este tiene 3 entradas : La pronunciación desconocida, el modelo HMM de las referencias o SMs. y el conocimiento léxico en términos de una red de estados finitos basada en un vocabulario de palabras. Como resultado del proceso se obtiene una celosía de hipótesis de palabras.

La tarea implica la comparación de un conjunto de parámetros extraídos de la señal de voz con una serie de modelos de unidades fonéticas. En nuestro caso, la unidad fonética básica de reconocimiento es la semisílaba, unidad elegida atendiendo al carácter silábico de la lengua castellana. Cada referencia o semisílaba es caracterizada por un modelo Oculto de Markov (HMM) y la media y varianza de la longitud de la semisílaba.

La señal de voz , por su parte, es filtrada (100 Hz-3400 Hz) y muestreada a 8 Khz. La señal digital es segmentada en tramas de 30 mseg. Se establece una parametrización LPC (Linear Predictive Coding) y una cuantificación vectorial (VQ).Cada trama de señal de voz viene representada por 3 símbolos enteros, que serán los componentes del vector de Observaciones O correspondiente a la señal de voz . Este vector será una entrada del algoritmo de "Spotting".

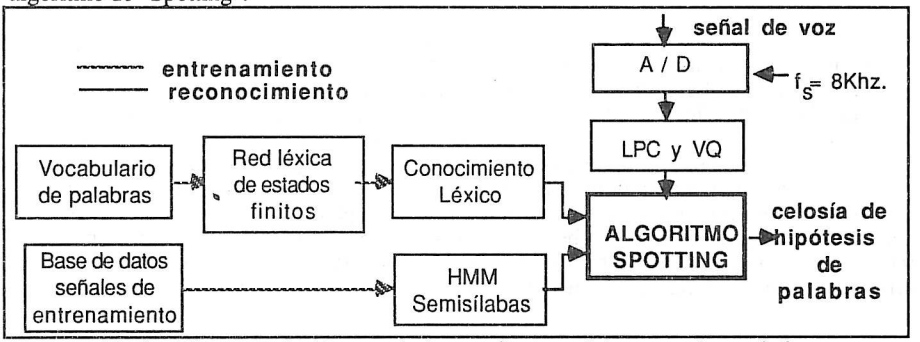

Fig.1 . Diagrama de bloques de la arquitectura del procesador acústico

3 . LOCALIZACION DE SEMISILABAS

El primer nivel de localización se establece sobre nuestra unidad básica : la semisílaba (SM). Todos los algoritmos implementados en este nivel tienen su base en el algoritmo de Viterbi, en consonancia con la teoría básica de los HMM.

El algoritmo de Viterbi toma como entradas la pronunciación desconocida , representada por el vector de observaciones O y el HMM de cada SM. o referencia. En él se desenvuelve un proceso que permite desarrollar caminos o secuencias de estados como resultado de establecer una comparación o "matching", basada en criterios de probabilidad, entre la referencia y partes de la pronunciación desconocida. El desarrollo del algoritmo puede ser implementado según una estructura en forma de enrejado o "trellis"; donde, según un proceso de izquierda a derecha , las probabilidades acumuladas asociadas a los caminos compiten para alcanzar cada nodo del "trellis". Sólo un único camino alcanza cada nodo.

La probabilidad acumulada está formada por las probabilidades de Observación **bi(Ot)** asociadas a cada estado y trama , y por las probabilidades de transición entre estados **aij** . La figura 2 muestra una competencia genérica entre caminos dentro del "trellis". **P(t,s)** corresponde a la probabilidad acumulada asociada a la trama **t-ésima** y estado **s,** como resultado de la competencia entre caminos en la transición entre los estados **s, s-1, s-2** en la trama **t-1-ésima** , y el estado s en la trama **t-ésima**. Todas las probabilidades son consideradas en forma logarítmica.

Para cada trama de entrada, se evalua la probabilidad de que cada unidad acabe en esa trama. Luego, para un proceso de Viterbi se obtiene un vector que contiene las puntuaciones

asociadas a esta evaluación. En este nivel de localización de SMs., se han desarrollado dos aproximaciones : El algoritmo de M PASOS y el de 1 PASO

3-1 ALGORITMO DE M PASOS

La filosofía "Word Spotting" implica la localización de unidades fonéticas en una frase . Estas unidades pueden estar situadas arbitrariamente en la señal de voz. Por tanto, los caminos que se desarrollen en el proceso de "matching" deben ser entendidos como potencialmente iniciables en cualquier trama . En este punto el algoritmo de Viterbi clásico establece la inicialización del proceso exclusivamente en la primera trama . Luego, extendiendo este principio a cualquier trama , se ha configurado el algoritmo de localización de M PASOS.

En este algoritmo, el punto de inicio del paso m-ésimo se sitúa en la trama m-ésima; o sea, ejecuta tantos procesos de Viterbi como tramas tenga la señal de entrada . Por tanto, todas los posibles caminos son evaluados. En este sentido, puede ser calificado como exacto. Sin embargo, esta evaluación implica un excesivo coste computacional.

3-2 ALGORITMO DE 1 PASO

Este algoritmo surge con la idea de reducir el coste computacional del anterior sin disminuir sus buenas prestaciones y a partir de la siguiente idea : En un único proceso de Viterbi es posible obtener los caminos de máxima probabilidad. Luego, cada trama de entrada puede ser punto de inicio de camino en el proceso de Viterbi; o sea, se relaja la restricción de inicio de camino del algoritmo de Viterbi . En este sentido, ha sido preciso adoptar un criterio de competencia en el primer estado. Por otra parte, debido a la desigual longitud de caminos, sus probabilidades asociadas deben ser normalizadas por la longitud del camino, para ,así, poder comparar cada camino en cada instante.

Los resultados obtenidos por el algoritmo de 1 PASO son prácticamente idénticos a los obtenidos por el algoritmo de M PASOS, pero el ahorro computacional es extraordinario.

4 - LOCALIZACION DE PALABRAS

El segundo nivel de localización en el proceso de "spotting" se establece sobre la palabra . Tal como ya se ha dicho, "Word Spotting" implica la localización de un pequeño vocabulario de palabras extraíbles de una conversación arbitraria .

El conocimiento léxico se caracteriza a partir de un diccionario en forma de árbol que contiene todas las pronunciaciones de las palabras en términos de semisílabas. A partir de él se genera una representación compacta basada en una gramática de estados finitos.

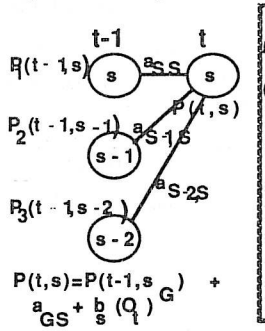

Fig.2 Competencia de caminos.

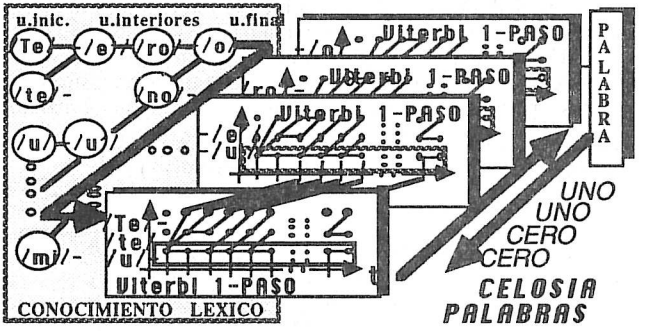

Fig 3.Esquema del proceso de Word Spotting

Para localizar una palabra en la señal de voz , el proceso de "spotting" debe partir de una unidad inicial y llegar a una unidad final de la cadena asociada . En este sentido, para

cada semisílaba de la cadena correspondiente a cada palabra, se desarrolla un proceso de Viterbi de 1 PASO. Por tanto, será necesario definir las transiciones entre unidades (SMs.).Estas son determinadas por el conocimiento léxico , el cual dirige el proceso de "spotting". Cuando el proceso alcanza el último estado en una unidad inicial o interior de una cadena , el conocimiento léxico y la probabilidad de duración de cada semisílaba conducen al algoritmo hacia la siguiente unidad léxica (SM), tal como indica la figura 3. El algoritmo ha sido modificado convenientemente para generar múltiples hipótesis en estas transiciones entre unidades. Esta modificación se concreta en considerar las N mejores secuencias de unidades léxicas en cada transición.

Finalmente, el algoritmo de "word spotting" proporciona para cada trama de entrada , en el último estado de cada proceso asociado a las semisílabas finales , la probabilidad de que cada palabra del vocabulario acabe en esa trama , así como los límites de las localizaciones . Toda esta información es incorporada a una celosía de palabras.

5-EXPERIMENTOS Y RESULTADOS

El procesador acústico ha sido testeado a partir de una aplicación específica : El reconocimiento de números castellanos pronunciados en un contexto de habla contínua.
Para ello se han utilizado 3 bases de datos de locutores:
DB1 : 10 locutores (5 hombres,5 mujeres). Compuesta por cadenas de enteros.
DB2 : 9 locutores (5 hombres,4 mujeres) .Compuesta por números telefónicos.
DB3 : 10 locutores (6 hombres,4 mujeres) . Compuesta por enteros del 0 al 1000.
DB1 ha sido la base utilizada para el entrenamiento, mientras que DB2 y DB3 lo han sido para el reconocimiento. Estos dos conjuntos de señales (DB1 y DB2-DB3) están caracterizados por diferentes velocidades de articulación, diferentes frases y diferentes locutores , de acuerdo con el contexto de de independencia del locutor de este trabajo.

El número de semisílabas para esta aplicación es de 67 y el número de palabras 32. El número de cadenas asociadas a las palabras es de 82.

La figura 4 muestra como se han obtenido y evaluado los resultados . En este caso se representan las mejores hipótesis (previamente ordenadas según máxima probabilidad). El eje vertical representa probabilidades logarítmicas normalizadas asociadas a cada hipótesis. El eje horizontal es un eje temporal (tramas en que ha sido dividida la señal). Los segmentos asociados a cada hipótesis aparecen colocados en el gráfico acorde con su localización en la señal de voz .
Para cada palabra, el proceso genera la siguiente información : Su localización óptima (de máxima probabilidad), su probabilidad asociada y los límites o fronteras de la localización (tramas de inicio y fin). Así, el primer segmento de la figura representa la primera hipótesis que corresponde a la palabra "cien", la cual es localizada entre las tramas 26 y 39 con máxima probabilidad (-3.23); y, si bien la trama inicial aparece como invariable, la trama final podría variar a largo de un intervalo entre las tramas 36 y 46.

Un analizador muy simple proporcionaría fácilmente la cadena de palabras componentes de la frase pronunciada . Esta tarea consistiría en combinar las hipótesis de palabras a partir de dos criterios : Consistencia temporal y máxima probabilidad.

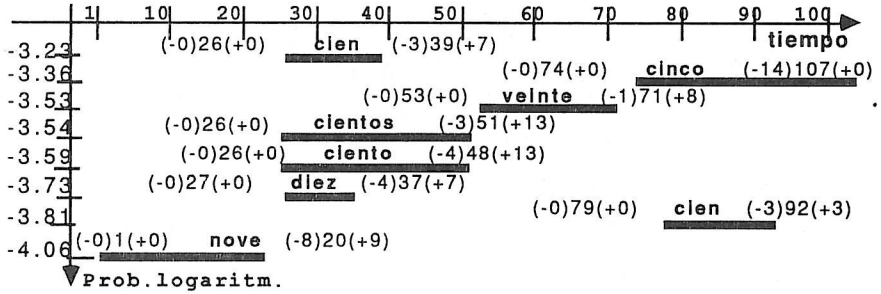

Fig. 4 . Parte significativa del resultado del proceso de "Word Spotting"

Se han realizado 2 tipos de experimentos : En el primero, se elige una sola hipótesis en las transiciones entre unidades (1 HIP.) y las demás son rechazadas . En el segundo, se consideran N hipótesis (MULTIPLES HIP., N=4)

Sobre la celosía de palabras obtenida, se definen un conjunto de niveles de hipótesis según el siguiente criterio : Un nivel de hipótesis se define como la posición que ocupa la palabra correcta en su posición correspondiente en la frase pronunciada, una vez las hipótesis han sido ordenadas según criterio de máxima probabilidad.

La figura 5a corresponde a las tasas de reconocimiento obtenidas para la base DB2, números telefónicos. La figura 5b corresponde a las obtenidas para la base DB3 o números enteros. En el primer experimento las tasas oscilaron entre el 74% para el primer nivel de hipótesis y el 93% para 5 niveles de hipótesis , para el caso de 1 HIP.; y fueron del 98% para el caso de 5 niveles de hipótesis y MULTIPLES HIP. En el segundo experimento, las tasas fueron del 82% y 95% para 1 HIP. y del 99% para MULTIPLES HIP. En este experimento se optimizan las tasas. Ello es debido al diferente número medio de palabras por frase en los dos casos (6.2 y 2.56) . Cuanto mayor sea este número, mayor coarticulación se produce.

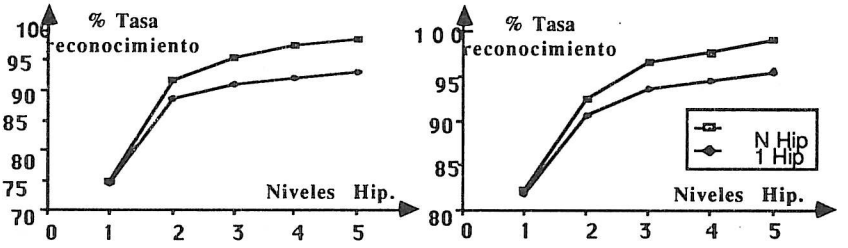

Fig. 5 Tasas de reconocimiento correspondientes a : a) Base DB2, números telefónicos b) Base DB3, números enteros del 0 al 1000.

6-CONCLUSIONES

Se ha implementado un procesador acústico para la lengua castellana basado en la filosofía "Word Spotting". El objetivo del trabajo era localizar palabras o unidades de reconocimiento en una frase. Para ello se han utilizado modelos HMM y se ha partido del algoritmo de Viterbi. Los resultados obtenidos en un contexto de habla contínua e independencia del locutor demuestran la eficiencia del algoritmo de "spotting" y las altas prestaciones del procesador acústico desarrollado, confirmando a la semisílaba como unidad robusta para el procesado acústico y a la palabra como apropiada para el procesado lingüístico.

REFERENCIAS

[1] J. Salavedra . " Processador Acústic de paraules basat en la Tècnica de "Word Spotting" pel reconeixement de la Parla Contínua". Projecte Fi de Carrera 1990.

[2] Lawrence R. Rabiner ."A tutorial on Hidden Markov Models and selected applications in speech recognition".Proceedings of the IEEE,vol.77, NO.2,February 1989 (pp.257-285)

[3] J.B. Mariño, C. Nadeu, E. Lleida, "Finite State Grammar Inference for Connected Word Recognition", EUSIPCO-88, 1059-1062, GRENOBLE-88.

[4] J.B. Mariño , E. Monte . "Generation of multiple hypothesis in connected phonetic-unit recognition by a modified one-stage dynamic programming algorithm". EUROSPEECH-89 . 408-411, Paris 1989